

Real-Time Hand Shape Classification

Jakub Nalepa^{1,2} and Michal Kawulok¹

¹ Institute of Informatics, Silesian University of Technology
Akademicka 16, 44-100 Gliwice, Poland
{jakub.nalepa,michal.kawulok}@polsl.pl

² Future Processing, Gliwice, Poland
jnalepa@future-processing.com

Abstract. The problem of hand shape classification is challenging since a hand is characterized by a large number of degrees of freedom. Numerous shape descriptors have been proposed and applied over the years to estimate and classify hand poses in reasonable time. In this paper we discuss our parallel framework for real-time hand shape classification applicable in real-time applications. We show how the number of gallery images influences the classification accuracy and execution time of the parallel algorithm. We present the speedup and efficiency analyses that prove the efficacy of the parallel implementation. Noteworthy, different methods can be used at each step of our parallel framework. Here, we combine the shape contexts with the appearance-based techniques to enhance the robustness of the algorithm and to increase the classification score. An extensive experimental study proves the superiority of the proposed approach over existing state-of-the-art methods.

1 Introduction

Hand gestures constitute an important source of non-verbal communication, either complementary to the speech, or the primary one for people with disabilities. The problem of hand gesture recognition has been given a considerable research attention due to a wide range of its practical applications, including human-computer interfaces [1, 2], virtual reality [3], telemedicine [4], videoconferencing [5], and more [6, 7]. The proposed approaches can be divided into hardware- and vision- based methods. The former utilize sensors, markers and other equipment to deliver an accurate gesture recognition, but they lack naturalness and are of a high cost. Vision-based methods are contact-free, but require designing advanced image analysis algorithms for robust classification. Thus, an additional effort is needed for applying these techniques in real-time applications.

Numerous algorithms for hand shape classification have emerged over the years. In the contour-based approaches, the shape boundary of a detected hand is considered to represent its geometric features. These methods include, among others, a very time-consuming approach based on the shape contexts analysis and estimating similarity between shapes [8], recently optimized by reducing the search space by using the mean distances and standard deviations of shape

contexts [9], and Hausdorff distance-based methods [10]. The main drawback of these contour-based approaches lies in their limited use in case of missing contour information.

In the appearance-based methods, not only is the contour utilized for shape features extraction, but also the shape’s internal region is analyzed. For example, an input color or greyscale image is processed in the orientation histograms approach [11] or an entire hand mask can be fed as an input to various template matching and moment-based methods [12]. An interesting and thorough survey on vision-based hand pose estimation methods was published by Erol et al. [13].

In this paper we discuss a fast parallel algorithm for hand shape classification. We show how the parallelization affects the classification time, and makes it possible to apply for searching large hand gesture sets in reasonable time. We present the speedup and efficiency of the parallel algorithm for various numbers of threads. Moreover, we show that combining the shape contexts with the appearance-based methods results in increasing the final classification score.

The paper is organized as follows. The hand shape classification algorithm is described in detail in Section 2. The experimental study is reported in Section 3 along with the description of the database of hand gestures. The paper is concluded and the directions of our future works are highlighted in Section 4.

2 Parallel Hand Shape Classification Algorithm

In this section we describe our parallel algorithm (PA) for hand shape classification [14]. First, the input image I_i is subject to skin segmentation, only if necessary (Alg. 1, lines 2–5). This step is undertaken if the shape features are to be extracted from the skin map of I_i . There exist a number of robust skin detection and segmentation techniques [15–19]. A thorough survey on current state-of-the-art skin detection approaches has been published recently [20]. Then, the image, either the skin mask or the original one, is normalized (line 6). The normalization procedure is presented in Fig. 1. An input image (A) or the skin map (B) is rotated (C) based on the position of wrist points, so as the hand is oriented upwards. Pixels below the wrist line are discarded, the image is cropped and downscaled to the width w_M (D).

Once the image is normalized, hand shape features are calculated in parallel (Alg. 1, lines 7–9), and the input image I_i is compared with the gallery images, also in parallel (lines 10–12). Finally, the classification result is returned (line 13). Noteworthy, the classification procedure can be executed for a number of input images in parallel. Thus, larger databases of input hand images can be analyzed significantly faster than using a sequential approach.

In the shape features calculation and shape classification stages we utilized the following state-of-the-art methods: (1) shape contexts analysis (SC) [8], (2) template matching (TM), (3) Hausdorff distance analysis (HD) [10], (4) comparison of the orientation histograms (HoG) [11], (5) Hu moments analysis (HM) [21], and two approaches combining the SC with the appearance-based methods: SC combined with the distance transform (SCDT) and SC enhanced by the orien-

Algorithm 1 Parallel hand shape classification.

```
1: parfor  $I_i \leftarrow I_1$  to  $I_N$  do
2:    $SkinMapIsAnalyzed \leftarrow \text{CheckIfSkinMapIsAnalyzed}(I_i)$ ;
3:   if  $SkinMapIsAnalyzed$  then
4:     Detect skin and create skin map;
5:   end if
6:   Normalize image; ▷ See Fig. 1
7:   parfor  $H_j \leftarrow H_1$  to  $H_M$  do
8:     Calculate  $j$ -th hand shape feature;
9:   end parfor
10:  parfor  $G_j \leftarrow G_1$  to  $G_g$  do
11:    Compare  $I_i$  with  $j$ -th gallery image;
12:  end parfor
13:  return final hand shape classification;
14: end parfor
```

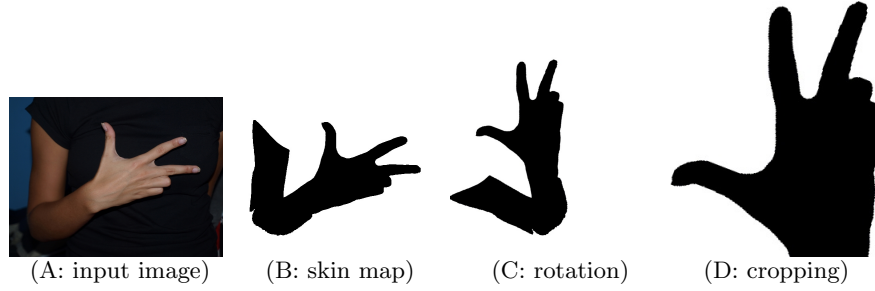


Fig. 1. Example of the hand skin map normalization.

tation histograms analysis (SCH). In the TM algorithm we used the following summation for comparing the overlapped patches of images I_1 and I_2 , of size $w_1 \times h_1$ and $w_2 \times h_2$, respectively:

$$R(i, j) = \sum_{a, b} (I_1(a, b) - I_2(i + a, j + b))^2, \quad (1)$$

where $0 \leq a \leq w_2 - 1$ and $0 \leq b \leq h_2 - 1$.

Let κ and λ be two contours compared in the SC method. For each point p_i^κ and p_i^λ , $i \in \{1, \dots, m\}$, where m is the number of contour points, belonging to κ and λ respectively, the coarse log-polar histogram h_i , i.e., the shape context, is calculated. It depicts the distribution of the remaining $(m - 1)$ points for each p_i . Let C_{ij} denote the cost of matching the points p_i^κ and p_j^λ , given as a chi-square distance between the corresponding shape contexts. Then, the total matching cost of two contours C is given as $C = \sum_i C(p_i^\kappa, p_{\pi(i)}^\lambda)$, where π is a permutation of the contour points. Clearly, the minimization of C is an instance of the bipartite matching problem. It can be solved in $O(n^3)$ time, where n is the number of sampled contour points, using the Hungarian method [9]. To speed

up the SC, we sample and analyze a subset of M_{SC} , $M_{SC} \ll m$, contour points of a shape. Additionally, the distance transform (DT) of the hand mask from the contour is performed. Given the DT, its histogram H_i is calculated for the image I_i . Then, the distance between the histograms H_1 and H_2 of two images I_1 and I_2 is found using the chi-square metric:

$$d(H_1, H_2) = \sum_B \frac{(H_1(B) - H_2(B))^2}{H_1(B) + H_2(B)}. \quad (2)$$

The final cost of shapes matching of the SCDT is given as: $C' = \alpha C + \beta d(H_1, H_2)$, where α and β are the weights. Values of C and $d(H_1, H_2)$ are normalized, therefore $0.0 \leq C \leq 1.0$ and $0.0 \leq d(H_1, H_2) \leq 1.0$. Similarly, the shape contexts are combined with the orientation histograms approach [11] using the same values of weights α and β .

3 Experimental Results

The PA was implemented in C++ language using the OpenMP interface. The experiments were conducted on a computer equipped with an Intel Xeon 3.2 GHz (16 GB RAM with 6 physical cores and 12 threads) processor having the following cache hierarchy: 6×32 kB of L1 instruction and data cache, 6×256 kB L2 cache and 12 MB of L3 cache. The settings used in both stages of the PA were tuned experimentally to the following values: $\alpha = 0.17$, $\beta = 1.0$, $w_M = 100$, $M_{SC} = 20$.

3.1 Database of Hand Gestures

The experimental study was carried out using our database of 499 color hand images of 15 gestures presented by 12 individuals³. Each gesture was presented n times, $27 \leq n \leq 39$. The images are associated with ground-truth binary masks indicating skin regions along with the ground-truth hand feature points. In this study, we omitted the skin segmentation and wrist localization stages, and used the ground-truth data for fair assessment of investigated techniques applied at other algorithm steps. Examples of ground-truth binary masks are presented in Fig. 2. Here, each gesture (1, 2, 3, 4, H, K, N, S) was presented by five individuals (I–V). It is easy to note that the difference between masks representing the same gesture (i.e. inner-class difference) may be significant, e.g. due to the hand rotation – see e.g., Fig. 2(N).

3.2 Classification Accuracy Analysis

The data set was split into a gallery (G) and a probe set (P) [22]. The gallery contains exactly g , $g \geq 1$, sample images per each gesture in the data set.

³ For more details see <http://sun.aei.polsl.pl/~jnalepa/BDAS2014>

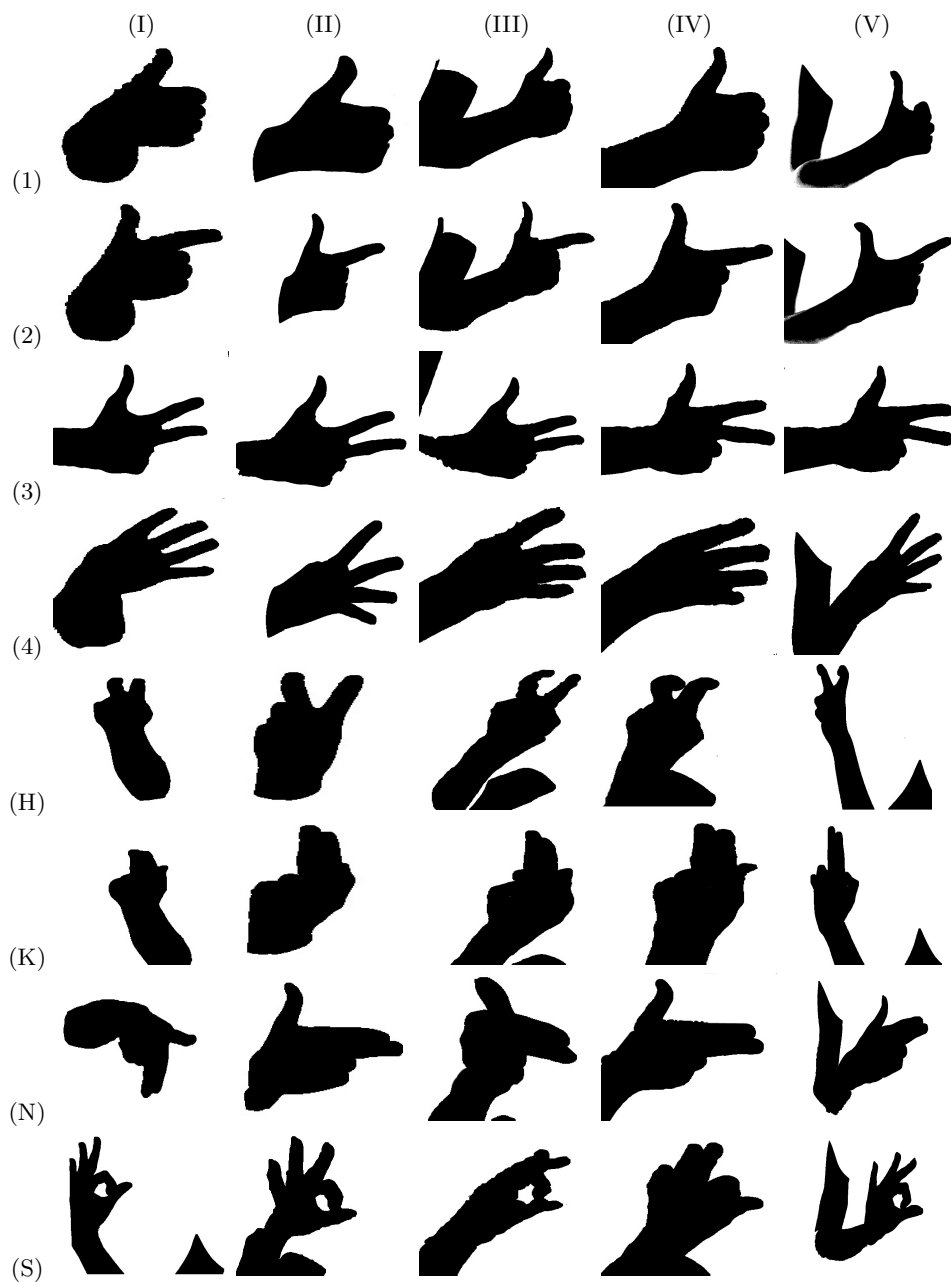


Fig. 2. Examples of ground-truth binary masks of various gestures (1, 2, 3, 4, H, K, N, S) presented by five individuals (I–V).

Then, the similarities of the images from P to those in G were found using the techniques outlined in Section 2. Classification effectiveness is assessed using its *rank* (R), $1 \leq R \leq |G|$. The rank is the position of the correct label on a list of gallery images sorted in the descending order by the similarity. If an image is classified correctly, then its rank is 1. The classification effectiveness for a given set is a percentage of correctly classified images. The analysis of the classification efficacy is performed based on cumulative response curves (CRCs).

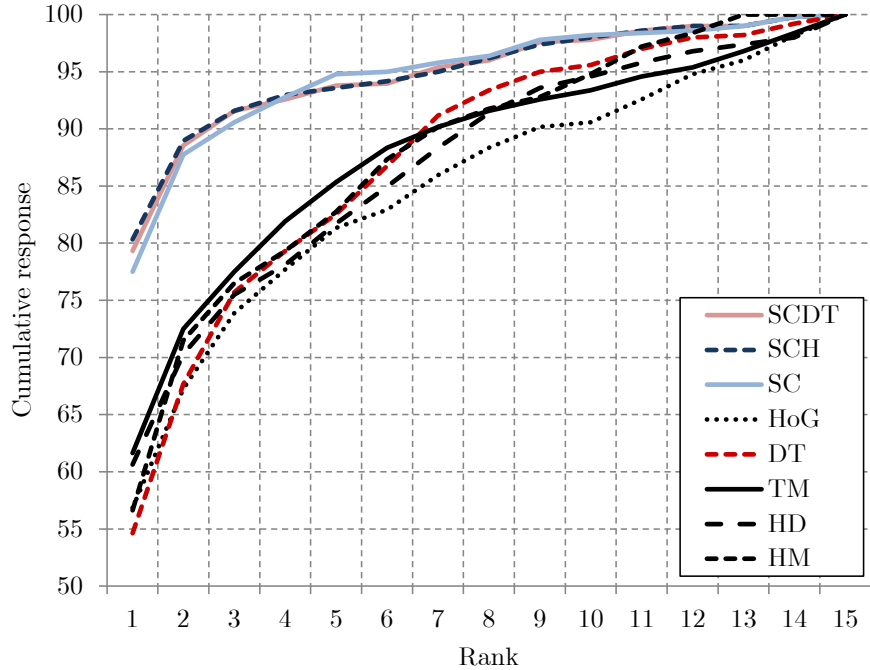


Fig. 3. Best CRCs for a single image per gesture in G .

The best CRCs for a single image per gesture in G are given in Fig. 3. We performed 27 classification experiments with no overlaps between the galleries. Then, the results were averaged and the best set of gallery images was determined. It is easy to see that the shape context methods, SCDT, SCH and SC, outperformed other techniques by at least ca. 15%, considering the correct classification (see rank 1). Additionally, the algorithms enhanced by the appearance-based approaches, SCDT and SCH outperformed standard SC by 2% and 3%.

Tab. 1 shows the average CR values for 4 initial ranks along with their standard deviations σ . In the average case for a single image per gesture in G (Tab. 1(A)), it is the SCDT method which turned out to be the best among the

Table 1. Average CR along with the standard deviation σ (best CR shown in boldface) for various number of gallery images g : (A) $g = 1$, (B) $g = 3$, (C) $g = 5$.

Method ↓	Rank 1 CR $\pm \sigma$	Rank 2 CR $\pm \sigma$	Rank 3 CR $\pm \sigma$	Rank 4 CR $\pm \sigma$
(A) SCDT	69.30 \pm 5.84	78.04 \pm 2.20	82.92 \pm 1.80	85.95 \pm 1.28
SCH	69.02 \pm 6.33	77.82 \pm 2.19	82.38 \pm 1.82	85.40 \pm 1.17
SC	69.04 \pm 5.75	77.97 \pm 2.27	82.63 \pm 1.88	85.82 \pm 1.05
HoG	44.14 \pm 7.64	57.56 \pm 3.05	65.25 \pm 1.49	70.37 \pm 1.52
DT	41.95 \pm 6.71	56.87 \pm 2.76	66.34 \pm 1.86	72.95 \pm 1.47
TM	51.38 \pm 7.16	63.69 \pm 2.40	69.48 \pm 1.82	74.05 \pm 1.32
HD	49.10 \pm 7.44	59.81 \pm 2.01	66.39 \pm 1.62	71.21 \pm 1.73
HM	46.69 \pm 7.07	60.86 \pm 3.48	68.46 \pm 2.52	74.57 \pm 1.62
(B) SCDT	76.66 \pm 2.84	82.17 \pm 1.43	85.71 \pm 0.90	88.47 \pm 1.19
SCH	77.10 \pm 1.56	82.11 \pm 0.87	86.23 \pm 0.61	88.07 \pm 0.70
SC	75.74 \pm 2.45	81.00 \pm 1.47	84.23 \pm 0.88	86.75 \pm 0.73
(C) SCDT	81.18 \pm 1.25	85.97 \pm 0.87	88.35 \pm 0.74	90.33 \pm 0.64
SCH	80.59 \pm 2.30	86.06 \pm 1.38	88.28 \pm 0.84	89.80 \pm 0.65
SC	80.17 \pm 1.69	85.18 \pm 1.57	87.63 \pm 0.86	89.21 \pm 0.55

investigated techniques, resulting in the highest CR values for each rank. Clearly, the choice of the image to G has a strong impact on the later classification score, and selecting more distinctive images significantly affects the final results (see σ in Tab 1(A)). Although the standard deviation of the rank 1 is the smallest for the shape context based algorithms (SCDT, SCH, SC), it is still noticeable and proves the methods to be quite sensitive to the choice of the gallery images.

Fig. 4 presents the CRCs for three ($g = 3$) most discriminative images, i.e. these that gave the best score for $g = 1$ for each method, per gesture in G . Providing multiple gallery entries improved the correct results in the initial ranks by at least 6% (SCDT, SCH and DT), up to 12% for the HD (see Fig. 3 and Fig. 4). On the one hand, the appearance-based methods (HoG and DT) performed poorly for both $g = 1$ and $g = 3$. On the other hand, combining them with the contour-based shape contexts technique resulted in the best responses. Therefore, these methods are complementary. Noteworthy, combining the SC with other contour-based methods did not improve the classification score. Tab. 1(B) and Tab. 1(C) presents the average (out of 20 experiments) CR and its corresponding σ for $g = 3$ and $g = 5$ for SC-based methods. The enhanced approaches outperformed the SC significantly. Moreover, adding new images to the gallery (i.e. increasing g) made the algorithms more independent from the choice of gallery images (the σ values dropped).

3.3 Speedup and Efficiency Analysis

In order to assess the performance of the PA, we measured the analysis time $\tau(1)$ of the sequential algorithm and of the PA, $\tau(T)$, for various numbers of threads T , and calculated the speedup $\mathcal{S} = \tau(1)/\tau(T)$, along with the efficiency

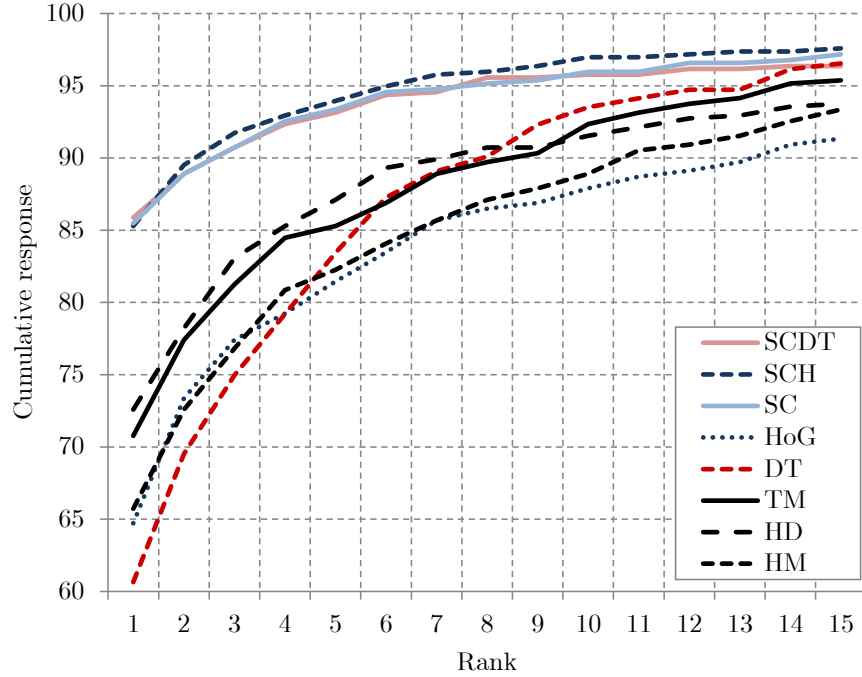


Fig. 4. Best CRCs for three images per gesture in G .

$E = S/T$ [23,24]. The analysis time τ consists of the feature extraction time τ_F and classification time τ_C , thus $\tau = \tau_F + \tau_C$.

We investigated the execution time of the sequential algorithm and the PA for both $g = 1$ and $g = 3$ for each technique. Also, we measured the analysis time for $g = 5$ for the best approaches, i.e. giving the best CR for a smaller number of images per gesture in G (SCDT, SCH and SC). In the latter case, we run the PA 20 times, with 5 random images representing a given gesture, using each mentioned approach. The average analysis time τ , along with the speedup S and efficiency E of the PA for various number of parallel threads T , are shown in Tab. 2. The HD is the most time-consuming classification approach. Although we significantly reduced the number of contour points considered in the SC, SCDT and SCH techniques, their sequential analysis time is still relatively high. The HM, HoG and DT turned out to be very fast for both $g = 1$ and $g = 3$. Providing new images to G increased the sequential analysis time of the TM and HD algorithms significantly.

The experiments performed for various number of threads T showed that the sequential algorithm can be speeded up almost linearly in case of more computationally intensive approaches. It is worth to mention that we experienced the superlinear speedup [25], i.e. $S > T$ and $E > 1.0$, while executing the PA

Table 2. Average analysis time τ , speedup \mathcal{S} and efficiency E of the PA for various numbers of threads T and gallery images g : (A) $g = 1$, (B) $g = 3$, (C) $g = 5$.

Method ↓	$T = 1$				$T = 2$				$T = 4$				$T = 8$			
	τ	τ	\mathcal{S}	E	τ	τ	\mathcal{S}	E	τ	τ	\mathcal{S}	E	τ	τ	\mathcal{S}	E
(A)	SCDT	32	16	1.96	0.98	9	3.45	0.86	6	5.22	0.65					
	SCH	39	19	2.01	1.01	11	3.56	0.89	7	5.29	0.66					
	SC	32	16	1.96	0.98	9	3.47	0.87	6	5.24	0.66					
	HoG	6	3	1.76	0.88	2	3.08	0.77	1	4.68	0.59					
	DT	4	2	1.70	0.85	1	3.02	0.76	1	4.61	0.58					
	TM	22	12	1.80	0.9	7	3.28	0.82	4	5.11	0.64					
	HD	83	41	2.05	1.03	21	3.89	0.97	14	6.00	0.75					
	HM	6	3	1.88	0.94	2	3.39	0.85	1	4.84	0.61					
(B)	SCDT	38	19	1.99	1.00	11	3.43	0.86	7	5.46	0.68					
	SCH	45	22	2.01	1.01	12	3.61	0.90	9	5.17	0.65					
	SC	37	19	1.97	0.99	11	3.41	0.85	7	5.36	0.67					
	HoG	6	3	1.71	0.86	2	2.94	0.74	1	4.58	0.57					
	DT	4	2	1.69	0.85	1	3.33	0.83	1	5.55	0.69					
	TM	54	31	1.76	0.88	17	3.22	0.81	11	5.15	0.64					
	HD	218	115	1.90	0.95	64	3.42	0.86	39	5.59	0.70					
	HM	6	3	1.82	0.91	2	3.32	0.83	1	4.25	0.53					
(C)	SCDT	52	25	2.09	1.05	13	3.93	0.98	9	6.04	0.76					
	SCH	58	27	2.15	1.08	15	3.86	0.97	10	5.84	0.73					
	SC	52	26	1.96	0.98	13	3.84	0.96	9	5.89	0.74					

with the HD, SCDT and SCH on two parallel threads ($T = 2$). Our preliminary studies indicated the local core caches as the source of superlinearity, however this issue requires further investigation. Applying the PA allows for increasing the G with a very fast analysis time and analyzing larger hand gesture databases. Thus, a more accurate classification can be performed in real-time (at more than 100 frames per second rate) using the available processor resources, e.g. the execution time of the SCDT ($g = 5$, $T = 8$) is more than 3.5 times lower than for a single thread and $g = 1$ with a very significant increase in the classification score.

4 Conclusions and Future Work

In this paper we discussed our parallel algorithm for fast hand shape classification. Introducing the parallelism allowed for decreasing the execution time of the sequential algorithm significantly. Moreover, we showed that the classification score can be boosted without increasing the execution time if the available processor resources are utilized. We experienced the superlinear speedup, which indicates high efficacy of the parallelization. Furthermore, we presented how the selection of gallery images influences the classification score.

Our ongoing research includes increasing the classification accuracy of the proposed parallel algorithm. We consider using radial Chebyshev moments here

as they occurred to be very effective for image retrieval purposes [26]. Also, we plan to investigate the fusing schemes of contour-based and appearance-based techniques to enhance the final classification. Additionally, we aim at applying the proposed approach for searching the space of the parameters that control a 3D hand model [27].

References

1. Ul Haq, E., Pirzada, S., Baig, M., Shin, H.: New hand gesture recognition method for mouse operations. In: *Circuits and Systems (MWSCAS), 2011 IEEE 54th International Midwest Symposium on.* (2011) 1–4
2. Czupryna, M., Kawulok, M.: Real-time vision pointer interface. In: *ELMAR, 2012 Proceedings.* (2012) 49–52
3. Shen, Y., Ong, S.K., Nee, A.Y.C.: Vision-based hand interaction in augmented reality environment. *Int. J. Hum. Comput. Interaction* **27**(6) (2011) 523–544
4. Wachs, J., Stern, H., Edan, Y., Gillam, M., Feied, C., Smith, M., Handler, J.: A real-time hand gesture interface for medical visualization applications. In: *App. of Soft Comp. Volume 36.* Springer Berlin Heidelberg (2006) 153–162
5. MacLean, J., Pantofaru, C., Wood, L., Herpers, R., Derpanis, K., Topalovic, D., Tsotsos, J.: Fast hand gesture recognition for real-time teleconferencing applications. In: *Proc. IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems.* (2001) 133–140
6. Grzejszczak, T., Nalepa, J., Kawulok, M.: Real-time wrist localization in hand silhouettes. In Burduk, R., Jackowski, K., Kurzynski, M., Wozniak, M., Zolnierok, A., eds.: *Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013. Volume 226 of Advances in Intelligent Systems and Computing.* Springer International Publishing (2013) 439–449
7. Nalepa, J., Grzejszczak, T., Kawulok, M.: Wrist localization in color images for hand gesture recognition. In Gruca, A., Czachorski, T., Kozielski, S., eds.: *Man-Machine Interactions 3. Volume 242 of Advances in Intelligent Systems and Computing.* Springer International Publishing (2014) 79–86
8. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. *IEEE TPAMI* **24**(4) (2002) 509–522
9. Lin, C.C., Chang, C.T.: A fast shape context matching using indexing. In: *Proc. IEEE ICGEC.* (2011) 17–20
10. Huttenlocher, D., Klanderman, G., Rucklidge, W.: Comparing images using the hausdorff distance. *IEEE TPAMI* **15**(9) (1993) 850–863
11. Freeman, W.T., Roth, M.: Orientation histograms for hand gesture recognition. Technical report, MERL (1994)
12. Thippur, A., Ek, C.H., Kjellstrom, H.: Inferring hand pose: A comparative study of visual shape features. In: *Proc. IEEE FG.* (2013) 1–8
13. Erol, A., Bebis, G., Nicolescu, M., Boyle, R.D., Twombly, X.: Vision-based hand pose estimation: A review. *Comp. Vis. and Im. Underst.* **108**(1–2) (2007) 52–73
14. Nalepa, J., Kawulok, M.: Parallel hand shape classification. In: *Proc. IEEE ISM.* (2013) 401–402
15. Jones, M., Rehg, J.: Statistical color models with application to skin detection. *International J. of Computer Vis.* **46** (2002) 81–96
16. Kawulok, M.: Energy-based blob analysis for improving precision of skin segmentation. *Multimedia Tools and Applications* **49**(3) (2010) 463–481

17. Kawulok, M., Kawulok, J., Nalepa, J., Papiez, M.: Skin detection using spatial analysis with adaptive seed. In: Proc. IEEE ICIP. (2013) 3720–3724
18. Kawulok, M.: Fast propagation-based skin regions segmentation in color images. In: Proc. IEEE FG. (2013) 1–7
19. Kawulok, M., Kawulok, J., Nalepa, J.: Spatial-based skin detection using discriminative skin-presence features. Pattern Recognition Letters (0) (2013) – in press.
20. Kawulok, M., Nalepa, J., Kawulok, J.: Skin detection and segmentation in color images. In Celebi, M.E., Smolka, B., eds.: Advances in Low-Level Color Image Processing. Volume 11 of Lecture Notes in Computational Vision and Biomechanics. Springer Netherlands (2014) 329–366
21. Hu, M.K.: Visual pattern recognition by moment invariants. Inf. Theory, IRE Trans. on **8**(2) (1962) 179–187
22. Phillips, P., Wechsler, H., Huang, J., Rauss, P.: The FERET database and evaluation procedure for face recognition algorithms. Im. and Vis. Comp. J. **16**(5) (1998) 295–306
23. Nalepa, J., Czech, Z.J.: A parallel heuristic algorithm to solve the vehicle routing problem with time windows. Studia Informatica **33**(1) (2012) 91–106
24. Nalepa, J., Blocho, M., Czech, Z.J.: Co-operation schemes for the parallel memetic algorithm. In: Parallel Processing and Applied Mathematics. Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) in press.
25. Chapman, B., Jost, G., Pas, R.v.d.: Using OpenMP: Portable Shared Memory Parallel Programming. The MIT Press (2007)
26. Celebi, M., Aslandogan, Y.: A comparative study of three moment-based shape descriptors. In: Proc. IEEE ITCC. Volume 1. (2005) 788–793
27. Šarić, M.: Libhand: A library for hand articulation (2011) Version 0.9.